Application Note: Quick Reference & Implementation Matrix

Companion Note to MCOG Working Paper v1.0

Decoupling Reasoning from Execution: A Requirement 1-of-1 Control Plane for Autonomous Agents

Michael Fargano, **Grok** (xAI), **Gemini** (Google)

August 2026 — MCOG Application Note v1.0

1 Introduction & Overview

This Application Note provides a high-yield implementation matrix and technical cheatsheet for the Moral Covenant of Governance (MCOG) Working Paper v1.0. It maps the evolutionary mechanisms of modern AI context management, analyzes historical failure modes in Conversational Control Planes (CCPs), details the operational execution of the Requirement 1-of-1 Enclosure Stack, and outlines the 4-stage engineering path for production adoption.

2 Architectural Trajectory & Historical Failure Modes

2.1 Core AI Mechanisms

Mechanism	Primary Function & Technical Specs
Vector RAG	Continuous semantic context retrieval via high-dimensional proximity mapping (cosine similarity, KNN searches, dense floating-point embeddings).
System Prompts / Rails	Behavioral alignment and direct instruction steering inside model context via LLM prompt prefixing and system instruction blocks.
LLM-as-a-Judge	Self-monitoring and output safety validation using a secondary LLM call (in-loop reflection, text-to-text safety evaluation passes).
MCP & Schema Validators	Structural typing and interface definitions for external tool invocation (JSON-Schema, Pydantic data contracts, Model Context Protocol).
OPA / Cedar Engines	Fine-grained, deterministic access control and authorization policy evaluation (native JSON evaluation, sub-millisecond compiled boolean rules).

2.2 History of Failure: The Conversational Control Plane (CCP)

Mechanism	Native Failure Mode	CCP Compromise (In-Loop)	Operational Impact
Vector RAG	**Semantic Drift:** Continuous distance metrics fail to enforce boolean boundaries.	Injects noisy, unverified context blocks directly into active prompt window.	High context pollution, probabilistic hallucination, prompt injection risk.
System Prompts	**Non-Determinism:** Models treat safety instructions as loose context hints.	Continuously stacks dynamic guardrail strings to cover edge cases.	Control-Plane Entropy Explosion; token overhead bloat; zero hard guarantees.
LLM-as-a-Judge	**Livelock & Fail-Open Risk:** Evaluator shares probabilistic vulnerabilities.	Policy violations produce natural language retry prompts (*"Please fix schema error X"*).	High serial latency, k-factor retry loops, continuous token drain.
MCP & Schemas	**Degraded Integration:** Schemas operate inside context rather than at edge.	Parsing failures fed back to generative model as text error messages for re-formatting.	Wasted inference turns attempting to coax structured outputs from generative nodes.
OPA / Cedar	**Misplaced Feedback:** Native policy engines degraded by conversational wrapping.	Policy rejections trigger multi-second natural language retry cycles instead of hard drops.	Converts microsecond native rules into expensive, multi-turn conversational negotiations.

[0]© 2026 by Michael Fargano, Grok (xAI), and Gemini (Google). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

3 The Solution: MCOG Requirement 1-of-1 Enclosure Stack

3.1 Paradigm Shift

Dimension	Legacy Control Plane (CCP)	Requirement 1-of-1 Enclosure Stack
Safety Location	**In-Loop:** Safety rules sit on token path inside reasoning plane.	**Out-of-Band:** Isolated reference monitor sitting strictly at execution boundary.
Decision Function	Probabilistic text negotiation / coaxing ($f(x) \to$ text).	Pure deterministic binary gate ($f_R(a) \in \{0, 1\}$).
Failure Handling	Generative retry loops ($k \geq 1$), fail-open / fail-negotiable.	Immediate fail-closed drop (0), structured non-LLM orchestrator fallback.
Scaling Dynamics	System bloat: global prompt rules scale linearly with complexity.	Requisite Variety: data-plane variety scales per query; rule body R remains modest.

3.2 Positive Laws & Implementation Sub-Axioms

Every technical sub-axiom derives directly from the three foundational moral pillars of the Moral Covenant of Governance (MCOG), aligned with Working Paper v1.0 §4.3:

- **Rights → Relative Invariant Completeness ($\mathcal{I}$) & Single Instrumented Egress:** Declared rights within $\mathcal{I}$ are non-negotiable and cannot be traded off for system utility or performance. All candidate actions must cross exactly one instrumented boundary so that no payload governed by $\mathcal{I}$ can escape via unmonitored side channels (raw sockets, unmediated `stdio`).
- **Stewardship → Modest Rule Body (R) & Ephemeral Trees:** Prevents bureaucratic control-plane decay[1] by keeping the compiled rule set R modest and scoped strictly to declared hard boundaries. Dynamic tree assembly ensures context trees dissolve post-execution without leaving persistent taxonomy or ACL bloat.
- **Reciprocity → Fail-Closed Out-of-Band Engine:** Respects the operational commons and multi-tenant resource constraints. Policy evaluation executes outside the model window using non-LLM deterministic solvers (OPA, Cedar, SMT). A policy denial results in an immediate fail-closed drop (0), ensuring a violation never becomes another tenant-funded conversational coaxing turn.

3.3 Mitigation Stack Summary

Component	Mechanism	Operational Impact
Mitigation #1: Dynamic Trees	Assembles ephemeral structural trees (Vectorless RAG) per query Q; dissolves post-execution.	Replaces continuous cosine distance with typed relational predicates; zero persistent schema decay.
Mitigation #2: Entropy-Gating	Context tree traversal bounded by next-token predictive entropy ($\bar{H}$).	Halts tree walks when $\bar{H} \leq H_{\text{target}}$ inside a wrapper bounded by $D_{\max}$. Matches variety to task without token bloat.
Mitigation #3: Binary Firewall	Isolated reference monitor executing compiled native policy solvers ($f_R(a)$).	Enforces non-bypassable egress; completely decouples generative reasoning from system execution.

4 Core Formulas Reference

1. **Admittance Decision Gate:** $\text{Admit}(a) \iff f_R(a) = 1, \quad f_R : \mathcal{A} \to \{0, 1\}$
2. **Conditional Closing Theorem:** Sub-Axioms Hold $\wedge$ R complete relative to $\mathcal{I} \implies$ No admitted action violates $\mathcal{I}$
3. **Conversational Control Plane Cost Model:** $C_{\text{CCP}} = (1 + k)(T_{\text{task}} + T_{\text{rails}}) + j \cdot C_{\text{judge}} + C_{\text{maint}}$
4. **Requirement 1-of-1 Cost Model:** $C_{\text{1-of-1}} = T_{\text{tree}}(\tau) + C_{\text{gate}} + p_{\text{deny}} \cdot C_{\text{fb}} \approx \alpha(\tau) \cdot T_{\text{task}} \quad (0 < \alpha(\tau) \leq 1)$

[1] Colloquially referred to as "Nanny Governance"; see companion paper *Costs of Nanny Governance v1.0.*

5 Definition of Variables

Variable / Symbol	Formal Definition	Operational Context / Description
$a \in \mathcal{A}$	Candidate Action Payload	The proposed state-changing operation or tool call generated by the data plane.
$f_R(a)$	Reference Monitor Function	Deterministic binary decision function returning strictly 1 (admit) or 0 (deny).
$\mathcal{I}$	Declared Invariants	Set of non-negotiable safety, privacy, and system boundaries.
R	Compiled Rule Body	Out-of-band boolean policy code (e.g., OPA Rego or Cedar ruleset).
Q	User Query	Natural language input parsed into typed relational entities and predicates.
$\bar{H}$	Generator Entropy	Mean next-token predictive entropy emitted during context traversal.
H_{target}	Target Entropy Threshold	Upper bound on acceptable uncertainty to trigger tree search termination.
$D_{\max}$	Maximum Depth Cap	Hard limit on structural context tree traversal depth to prevent infinite loops.
T_{task}	Task Context Tokens	Baseline context window token volume required for core task reasoning.
T_{rails}	Rail Overhead Tokens	Additional token volume consumed by in-prompt guardrails and system safety text.
k	Livelock Factor	Mean extra generation turns caused by in-loop policy negotiation and retry loops.
j	Judge Routing Fraction	Ratio of execution turns evaluated by secondary LLM safety models ($0 \leq j \leq 1$).
C_{judge}	Judge Evaluation Cost	Direct monetary cost incurred per secondary LLM verification call.
C_{maint}	Maintenance Overhead	Continuous operational cost of updating, debugging, and tuning prompt-rail strings.
$T_{\text{tree}}(\tau)$	Pruned Tree Tokens	Reduced token volume yielded by dynamic tree pruning at entropy threshold τ.
$\alpha(\tau)$	Context Retention Ratio	Token reduction factor achieved by lazy descent ($\alpha(\tau) = T_{\text{tree}}(\tau)/T_{\text{task}}$).
C_{gate}	Monitor Execution Cost	Computational cost of out-of-band evaluation ($C_{\text{gate}} \approx 0$ on LLM token ledger).
p_{deny}	Empirical Denial Rate	Observed probability of candidate action rejection at the binary egress firewall.
C_{fb}	Fallback Handler Cost	Non-LLM orchestrator cost executed upon action rejection (e.g., user prompt or log).

6 Four-Stage Implementation Matrix

Stage	Action / Architecture	Detailed Execution Steps	Exit Criteria / Done When
1	**Close Egress**	Encapsulate agent reasoning inside isolated sandbox (microVM / Wasm / container with eBPF hooks). Route all outgoing API, network, and system mutations through a single, explicit execution proxy.	Zero raw host network sockets or unmediated `stdio` channels available to model data plane.
2	**Compile $\mathcal{I} \to R$**	Declare $\mathcal{I}$ from system boundaries and compile directly to R. Translate invariants into compiled, non-LLM boolean policy solvers (OPA Rego or Cedar). Configure egress proxy to evaluate $f_R(a) \in \{0, 1\}$.	Reference monitor evaluates all egress actions out-of-band; defaults strictly to 0 (deny) on fault or parsing failure.
3	**Structured Errors Only**	Remove in-loop natural language retry prompts and LLM-as-a-judge re-evaluations. Denied actions return structured machine status codes to non-LLM orchestrator.	Retry factor drops to zero ($k \to 0$) on token path; no policy rejections passed back into model reasoning context.
4	**Instrument**	Deploy real-time telemetric counters for entropy threshold ($\bar{H}$), context retention ($\alpha(\tau)$), retry turns (k), judge calls (j), and denial rates (p_{deny}).	Telemetry active on at least one live production workflow; percentages published strictly from measured k, j, $\alpha(\tau)$, and p_{deny}.

7 List of Acronyms

Acronym	Definition
ACL	Access Control List
AIoT	Artificial Intelligence of Things
CCP	Conversational Control Plane
eBPF	Extended Berkeley Packet Filter
JSON	JavaScript Object Notation
KNN	k-Nearest Neighbors
LLM	Large Language Model
MCOG	Moral Covenant of Governance
MCP	Model Context Protocol
OPA	Open Policy Agent
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
SMT	Satisfiability Modulo Theories